DD2476 Search Engines and Information Retrieval Systems

Project 2: Tell Me a Story

Contact: Simon Stenström, Findwise (simon.stenstrom@findwise.se, 073-616 35 34, Sveavägen 28-30 (Hötorget))

This project is worth 3 ECTS credits. This means that it is expected to require 80 hours of work for each person in the group. The project formulation, method, and results are presented in a report as well as in a poster session. For more details, look at the course homepage, under Project in the menu.

Problem

Generating text from other texts, is an interesting information retrieval task. You need to extract content that can be combined to create new sentences. By making sure that the texts that you combine is about the same subject, and figuring out how sentences are built, the results can be inspiring. For this project, the meaning of the generated text does not matter. Instead we want to generate text that could have been created by a person.

Assignment

- Create a simple crawler for your use case
- Download the texts (together with some metadata about it)
- Extract sentences from the texts and store them in a search engine
- Combine the sentences, according to rules (depending on the metadata or sentence structure) or machine learning techniques
- Present the new stories as a web page

The text source can for example be news, song lyrics or children books.

Advanced techniques that can be used are:

- Part of speech tagging (for example https://github.com/EmilStenstrom/heroku-stagger for Swedish)
- Machine learning (for example https://en.wikipedia.org/wiki/Deep_learning)

The complexity of this project depends a lot on what you want to do. Feel free to play around to see what gives you the best results!

About Findwise

Findwise is a growing IT consultancy company, founded in 2005 by a team of experts from the enterprise search industry. The company currently employs about 100 people and has offices in Stockholm, Gothenburg, Copenhagen and Warsaw.

We aim to add business value for organizations where information is a priority by helping them to access, manage and use information. Findability by Findwise is all about creating search solutions that maximize business value gained from search technology investments. We create search solutions for intranets, web, e commerce and applications and make sure these are implemented to support and strengthen your business processes and help your organization reach business goals. We offer consulting and implementation of all leading platforms including Autonomy IDOL, Microsoft FAST ESP, Google GSA and the open source search engines Apache Solr and Elastic Search.

Findwise is fully customer-oriented. At the same time, our ambition is to be the best workplace in the industry, capable of attracting and retaining the best talents.